NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

1

An Experimental System for the Exchange of Scientific Information

M. Kochen and E. Wong
IBM Research Center

1. Introduction.

From the data collected in a study (1) on the journal reading habits of physicists and chemists, it can be estimated that if Journals of a given discipline (e.g. chemistry) are ranked according to decreasing frequency of being read, the probability p_i with which a journal of rank r is read varies approximately as a Yule distribution (2), i.e., $p_i \sim \frac{1}{r^\alpha}$, α being approximately one in this case. The salient characteristic of the Yule distribution is its long tail (slowly decreasing for increasing r). Thus, while 10 journals account for 50% of the reading, the remaining amount of reading is spread over a large number of different journals. While the few most frequently read journals are probably read by everyone in the same general field of interest, the remaining journals differ a great deal from person to person depending ~~strongly on the~~ special ~~interest~~ of the reader. In the face of rapid increases in the number of published journals, mostly in areas of high specialization, it is becoming increasingly difficult for an individual to discover items of potential value without an enormously increased reading load. The situation is even worse for items in unexpected or unknown sources. The discovery of these "rare" items of interest is the problem to which the present system is addressed

Cont. on p.

cont'd from p. 1

To a certain extent the problem of finding rare items is automatically alleviated in practice by extensive information exchanges among scientists with similar interests. Similarity in interests leads to "cliques" or "clusters" within which channels for efficient information exchange exist. The primary goal of the system being considered here is to discover, formalize, and utilize these clusters for the purpose of increasing the likelihood of discovering a rare item of information for an individual.

→ Cont'd on p. 3

The effectiveness of such a system can be estimated from the following considerations: assume that n scientists of the same interest-cluster consult the infrequently read journals in a statistically independent manner. Let q be the probability of finding an item of interest in these unusual sources upon consulting them. Let p be the total probability of consulting such journals for each person. In a cluster where such a discovery by one member would mean automatic discovery by all the members, the probability of a member discovering a rare item is $1 - (1 - qp)^n$ or about $1 - e^{-npq}$, as compared with qp for an isolated individual.

In practice, even better performance can be expected. The probability of discovery on an individual basis, i.e., pq, varies from person to person. There usually exists one or more members in any cluster who is much better informed than others. The system of exchange within a cluster has the effect of raising every member to at least the same degree of informedness as that of

Cont'd from p. 2

~~the best informed member.~~

Compared to a system where information is disseminated from a central source, this system of exchange through clustering has several distinct attractive features. First, since items distributed are "rare" items, the amount of information of minor interest or no interest at all to a participant is kept to a minimum. Secondly, since this is a system of exchange, the task of administering such a program is kept simple. This

~~The remainder of the~~ paper describes the design of the system, formation of clusters, and analysis of some preliminary performance data.

2. The Experimental System.

In the experimental system to be described here, people are grouped according to the similarities of their reading interests. This is, of course, only one of several possible relevant criteria for grouping interests besides what is read: what courses were best liked, what papers were written, the responses to keywords, etc. Reading was chosen here because it was particularly simple to obtain data,for, using a variant of a procedure used by King and Tanimoto.* (3)

The library, acting as a central message exchange, was supplied with a distribution list for each participant, listing all the other participants with

* They presented 20 respondents with the Table of Contents of a few selected journals, and asked them to check those they would read.

very similar reading interests. Any participant can inject items of information into the system by submitting to the library by telephone or in writing. The item might be: a complete or partial description of a particularly recommendable article which others are not likely to have come across; a technical question on which help from someone who might be uniquely qualified to help, is solicited; an idea for an experiment, a device, or a theoretical study on which reactions or comments are desired; a new finding to be announced to those interested; etc. If the originator submits it in writing, he records it on a special card, which has no rigid format except that it classifies the nature of the entry, and sends it to the library. If he telephones, the library prepares this card.

On receipt of such an item, the library duplicates this as many times as there are members on the sender's distribution list, and disseminates it to them. In order to monitor the recipients' responses for experimental purposes, a recipient receives, together with the duplicated entry, a simple response card, similar to that used in the SDI system. (4) On this, he indicates whether the item of information he received was of interest and/or new, and he sends the card for analysis.

This system shares with the SDI (Selective Dissemination of Information) the feature of trying to supply information about the scientific literature which would otherwise not be readily available to participants. Because the SDI system depends on a central source for scanning, selecting, and abstracting the literature, the amount and quality of the service depends only on this source, not on the number or level of the participants. In the system described here, the latter situation is the case, and, as pointed out above, the quality of service can be made very high by narrowly restricting membership in an interest cluster and increasing the number of participants.

In the future, the grouping of participants will have to be revised and checked periodically to take into account shifting interests, additional participants, quality and quantity of service. The data for this arises from the responses which recipients of information feed back into the system on a continuing basis. The "system" thus has two major functions, which may eventually be automatic:

(1) transmission, duplication and routing of information;

(2) continually sensing the state of the system and using this information to control its growth and operation.

3. Design of the Experiment.

To nucleate the system, it was decided to solicit participation from only certain members of IBM Research who were willing to take considerable initiative in contributing to its success. On the basis of a letter which was circulated to over 100 staff members of IBM Research, explaining the proposed system, thirty volunteered to participate, as a start. This group should not be regarded as a sample from which to draw substantive conclusions about the applicability to other groups, but this was not our goal. The purpose of this study was to demonstrate that there exists a system which would enable its members to obtain greater access to the literature with relatively little effort. To test the professional reading interests of these respondents (the network of people and set of procedures), the following crude method was used while an improved testing procedure is under development.

A random sample of 200 articles, represented by title and author only, was selected from the winter 1961 issues of about 450 English-language technical journals available in the library of the IBM Research Center. Each of the 30 respondents was asked to indicate, on a four-point ordered scale, to what extent he would be interested in the article on the basis of title-author. This test was administered through interviewing, along with a number of open-response questions designed to further characterize the respondents' professional interests, usage of the literature, and information needs.

In the analysis to be described, the responses were grouped into two categories, distinguishing no interest at all from its opposite: that is, the three response categories indicating degrees of positive interest were lumped together. This was done only to keep the consequent computations within reasonable bounds. The data was summarised in a table listing the respondents as row headings, and the articles in the sample as column headings. If a particular respondent expressed interest in a specific article, a 1 was entered in the cell corresponding to the appropriate row and column; for no interest, nothing was recorded, and it was treated as a 0 entry. Had the four-point scale been used, each article would be allowed three columns, representing the three categories of positive interest; a 0 or 1 would again be entered into the appropriate cell, and the analysis would proceed exactly as described, except that a 30 x 600 rather than a 30 x 200 table must be dealt with.

Inasmuch as the procedures for testing, sampling and validating statistical inferences are still under development, the detailed methodological considerations will be deferred to a later paper.

4. Clustering Analysis.

(a) Measure of Similarity:

Let r_{ik} be the response of the i^{th} person to the k^{th} article such that

$$r_{ik} \begin{array}{l} = \\ = \end{array} \left\{ \begin{array}{lll} 1, & \text{if interested,} & i = 1, \ldots, n \\ 0, & \text{if not,} & j = 1, \ldots, N. \end{array} \right. \qquad (1)$$

Here, n is the number of people (30 at the time of this report) and N the number of articles used to test them (200 in the first trial). Let R be the matrix with elements r_{ik}. Define the matrix C

$$C = R\overline{R} , \qquad (2)$$

where $\overline{R}$ denotes the transpose of R . A typical element c_{ij} of C represents the number of articles in which the interests of i and j co-occur.

The similarity, or association factor, s_{ij} between two people i and j is defined as

$$s_{ij} \triangleq \frac{N \times c_{ij}}{c_{ii} \quad c_{jj}} \qquad (3)$$

A number of other measures of association has been used, and five of these are listed in Table 1 for comparison.

The definition of Stiles (5) is a form of the chi-square formula on a 2 x 2 contingency table and includes the Yates' correction.(6)

King and Tanimoto (3) used two different measures. The first of these s_{ij} is called the similarity measure, the second, d_{ij} , distance measure, which is simply the negative log of s_{ij} . Our own definition was derived from the following considerations.

Author	Association Factor
Stiles (5)	$s_{ij} = \log_{10} \dfrac{N\left(\left| c_{ij}N - c_{ii}c_{jj} \right| - \frac{N}{2}\right)^2}{c_{ii}c_{jj}\,(N - c_{ii})\,(N - c_{jj})}$
Baxendale (7)	$s_{ij} = \dfrac{c_{ij}}{N}$
King-Tanimoto (3)	$s_{ij} = \dfrac{c_{ij}}{(c_{ii} + c_{jj} - c_{ij})}$, $d_{ij} = -\log s_{ij}$
Luhn-Savage* (4)	$s_{ij} = \dfrac{c_{ij}}{c_{jj}}$
Kochen-Wong	$s_{ij} = \dfrac{N\,c_{ij}}{c_{ii}\;c_{jj}}$

Table 1 - A Comparison of Association Factors

Assume that person i responds favorably to an article with probability p_i , and that responses to successive articles are independent (the independence of successive responses is a problem involved in sampling). If the frequency ratios $\frac{c_{ii}}{N}$ are taken to be the estimates of p_i , the mean number of coincidences between i and j is given by:

* Unlike the other measures, the Luhn-Savage definition measures association between different entities (documents and people), and is not symmetric, i.e., $s_{ij} \neq s_{ji}$.

$$\mu_{ij} = \frac{c_{ii}\, c_{jj}}{N} \tag{4}$$

Our measure of association is now defined as the ratio of the actual coincidence c_{ij} over the coincidence expected on the basis of independence,

$$s_{ij} = \frac{N\, c_{jj}}{c_{ii}\, c_{jj}} \tag{5}$$

In addition to computational ease, this definition has the advantage of possessing a simple intuitive interpretation. For example, for $s_{ij} = 5$ one can say that the actual coincidence between i and j is five times what is expected on the basis of zero-association (independence). Values of s_{ij} greater than 1 indicate positive association, $s_{ij} = 1$, zero association, and $s_{ij} < 1$ negative association, or dissociation. (The quantity $\log s_{ij}$ reflects these properties directly, but distorts the scale in an undesirable manner.) When the cluster finding procedure is fully programmed, it would be desirable to use a statistically more satisfactory definition, like

$$s_{ij} = \frac{\left[N\, c_{ij} - c_{ii}\, c_{jj}\right]^{2}}{c_{ii}\, c_{jj}\, (N - c_{ii})\, (N - c_{jj})} \; . \tag{6}$$

This is closely related to that of Stiles, and is based on the 2 x 2 contingency table shown in Table 2.

			Person i		
			interested	not interested	total
Person j	interested	Yes	c_{ij}	$c_{jj} - c_{ij}$	c_{jj}
		No	$c_{ii} - c_{ij}$	$N - c_{ij} - c_{ii} - c_{jj}$	$N - c_{jj}$
total			c_{ii}	$N - c_{ii}$	N

Table 2

A particular advantage of this definition is that it can be extended easily to deal with multiple-point scale response, e.g. (not interested, interested, very interested, and vital). In such cases one merely has to expand the contingency table and use a general version of this formula as a measure of association (8).

(b) Definition of Cluster:

A set C of k people is said to form a cluster relative to threshold ϵ if for every i, j in C,

$$s_{ij} \geq \epsilon . \qquad (7)$$

The quantity ϵ is a parameter to be chosen a priori, and determines the "strength" of the clusters formed. In general, an increase in ϵ will cause smaller and more "closely-knit" clusters to be formed.

The choice for a suitable definition of "cluster" poses a difficult problem which has occurred in a wide variety of applications (9, 10, 11, 12). Several definitions of "cluster" have been proposed (12, 13, 14, 15). The final choice adopted here was conservative in that it is required that in every "cluster" the association of reading interest between <u>any</u> two people must equal or exceed a certain minimum level. This choice of a "narrow cluster" definition was designed to insure that the clusters be homogeneous and closely-knit groups at the risk of leaving out people who may properly belong to clusters. The point to be emphasized here is that this concept of two individuals with a high degree of association may belong to different clusters by virtue of their associations with other people.

Thus the problem of finding clusters can be stated as follows:

Given a collection of people, find sets $C_1, C_2, \ldots, C_m$ (not necessarily disjoint), such that $s_{ij} \geq s$ for i and j belonging to the same C.

The clusters thus defined may not be unique. Furthermore, there exists no known schemes, other than exhaustive ones, for finding the clusters. Therefore, it is of considerable practical importance for an algorithm to be developed for forming the clusters to be derived. An algorithm based, in part, on the kind of heuristic devices used by people in extracting clusters from the data to accomplish this has been developed and will be described in the next section.

(c) Algorithm for Cluster-Formation:

The algorithm can best be described by an illustrative example. The association matrix for this example is shown in Figure 1, where the diagonal terms are omitted.

	1	2	3	4	5	6	7	8
1	X	6	2.4	6	10	0	2.2	7.8
2		X	4.2	4	10	4.3	5.7	3.4
3			X	1.4	4.2	4.3	3.8	0
4				X	6	2.8	2.5	4.5
5					X	4.3	3.8	3.4
6						X	3.8	0
7							X	3
8								X

Figure 1

In the procedure outlined below, a second paremeter, ϵ', $\epsilon' > \epsilon$ is used. For this example, $\epsilon = 3$ and $\epsilon' = 4$.

Procedure	Example
Step 1. For each i, starting with $i = 1$, find all the j's for which $s_{ij} \geq \epsilon'$	$i = 1,\ j = 2, 4, 5, 8$
Step 2. Form the set σ_i, containing as elements i and the j's found in step 1. Discard any set which is entirely contained in a previous set or contains less than 4 elements.	$\sigma_1 = (1, 2, 4, 5, 8)$

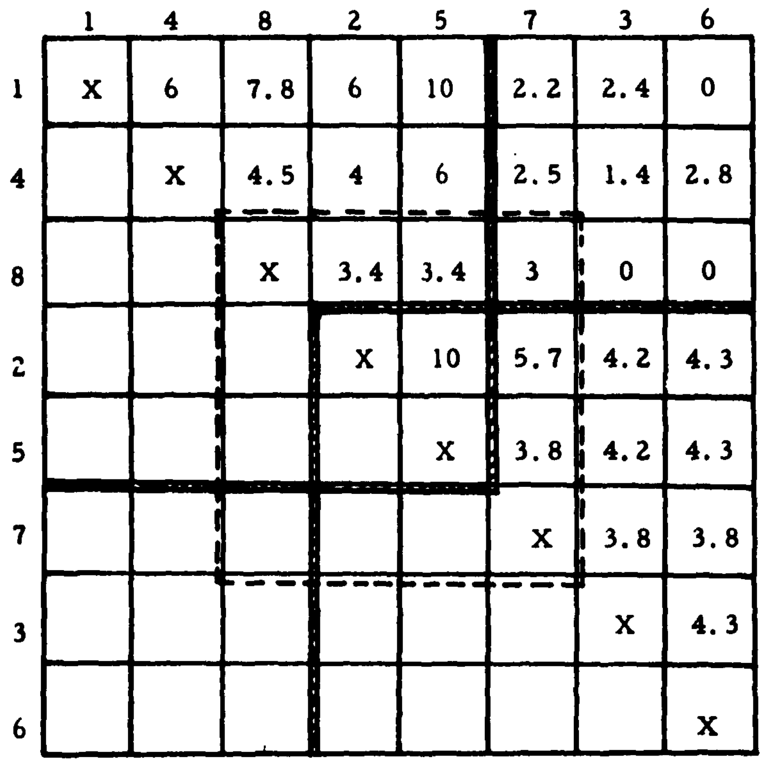
Application of Steps 1 and 2 to the numerical data of Figure 1 results in the following σ_i.

$$\sigma_1 = (1, 2, 4, 5, 8),$$
$$\sigma_2 = (1, 2, 3, 4, 5, 6, 7).$$

Procedure	Example
Step 3. Order the elements in each of the σ_i so that for every pair, j and k, in σ_i for which $s_{jk} < \epsilon$, j and k are on different sides of every ℓ i σ_i for which $s_{j\ell} \geq \epsilon$, and $s_{k\ell} \geq \epsilon$.	$\sigma_2 = (1, 2, 3, 4, 5, 6, 7)$. After ordering $\sigma_2 = (1, 4)(2, 5)(3, 6, 7)$, where elements within the same parenthesis may be reordered at will. Note that if $j = 1$, $k = 3$, $s_{ij} = 2.4$, so that $s_{13} < 3$; the only ℓ for which both $s_{1\ell}$ and $s_{3\ell}$ exceed 3 is $\ell = 2$ and $\ell = 5$; thus, both 2 and 5 must be placed between 1 and 3.

Procedure	Example
Step 4. Combine the partly ordered sets, and apply the requirement of step 3.	Combining σ_1 and σ_2 results in (8) (1, 4) (2, 5) (3, 6, 7). After applying the requirement of step 3, this becomes (1, 4) (8) (2, 5) (7) (3, 6).

The association matrix with rows and columns properly reorded is shown in Figure 2.



	1	4	8	2	5	7	3	6
1	X	6	7.8	6	10	2.2	2.4	0
4		X	4.5	4	6	2.5	1.4	2.8
8			X	3.4	3.4	3	0	0
2				X	10	5.7	4.2	4.3
5					X	3.8	4.2	4.3
7						X	3.8	3.8
3							X	4.3
6								X

Figure 2

The clusters and their interrelationships are apparent at a glance from Figure 2. If one adheres to the definition strictly, there are three clusters for this example. However, for operational purposes the two important clusters are (1, 4, 8, 2, 5) and (2, 5, 7, 3, 6) . It should be emphasized that the procedure outlined earlier does not merely find clusters. In fact, for pure enumeration of clusters there may well be more efficient procedures. In the process of finding the clusters, it has been possible to display the relationship among clusters in a succint manner.

The matrix-permutation procedure was based on several assumptions concerning the nature of the population (16) .

(1) Clustering to a large extent exists among the members of the population.

(2) The clusters are either isolated or overlap in a simple way. This assumption is equivalent to the hypothesis that the rows and columns of matrix S can be permuted to have a structure as shown in Figure 3 , where every entry in the shaded area (principal submatrices) is greater or equal to the threshold ϵ , and every entry in the unshaded area is less than ϵ .

This assumption implies that overlaps such as those shown in Figure 4 do not occur.

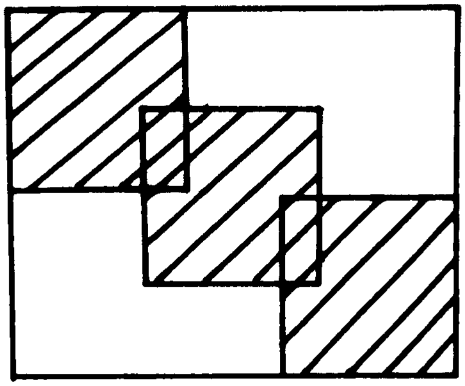


Figure 3

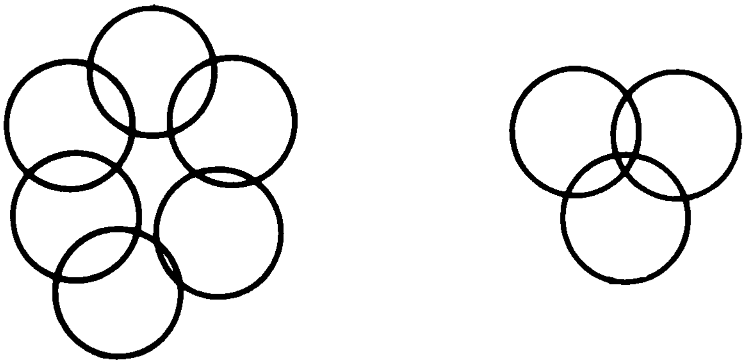


Figure 4

This assumption is only approximately valid in practice. That is, the resultant matrix will have structure like the one shown in Figure 3, but will have entries greater than ϵ in the unshaded areas.

There are two parameters in the procedure. One of these, ϵ , defines and controls the clusters that are obtained. The other parameters, ϵ', is used to initiate the procedure, and should not affect the final clusters that are found. A number of ways of choosing these paramaters, in a given problem are being investigated. One idea is to let ϵ' and ϵ be constant multiples of the average association, i.e.,

$$S = \frac{k_1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} s_{ij}, \qquad (8)$$

$$\epsilon = \frac{k_2}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} s_{ij}, \qquad (9)$$

The constants k_1 and k_2 are to be experimentally determined once for all, and would not vary from problem to problem.

5. Conclusion.

~~Of the 30 people tested, substantial clustering was found to exist for 15 people. These 15 people formed two clusters. It is of interest to note that the interests of members of both clusters are primarily in the physical sciences (physics, chemistry, metallurgy, etc.). The failure of the other participants to cluster is probably due to the insufficient number of people in each specialty.~~

During the initial 4 weeks, the system was in a testing phase. In order to obtain a substantial amount of data in a short period of time, normal operating procedures were deviated from in two important aspects. First, items were distributed to members of both clusters regardless of the source of the item. During normal operation the distribution will be confined to the cluster from which the item is initiated. Secondly, no attempt was made to confine the exchange to only rare items. At the end of the testing period, the participants were informed that only items of unusual interest and from unusual sources should be reported, in conformance with the primary aim of the system.

During the four-week testing period, a total of 41 items were initiated and distributed. Only one item failed to evoke any favorable response, i.e., interested. Of the 41 items, 35 were initiated by members of cluster # 1 and 6 by members of cluster # 2 . The acceptance rates are shown in Table 3.

Cluster	Average Percent of Acceptance per Person	
	Items initiated from within the cluster	Items initiated from without the cluster
# 1	18.5	4.1
# 2	46.5	5.1

Table 3.

The figures of 18.5% and 46.5% represent approximately improvements in acceptance rate of five-fold and nine-fold respectively. The improvement ratio can be compared with the average association factor for the two clusters, the association factor having precisely the interpretation of expected improvement ratio. The comparison is shown in Table 4.

Cluster	Average Association Factor	Improvement ratio
1	4.66	4.5
2	4.95	9.1

Table 4

The agreement for cluster #1 is obviously good. The deviation from agreement for cluster #2 is probably due to the small sample (6 items initiated from cluster #2).

Although the limited data obtained thus far does not admit general conclusions, the aim of effective discovery and dissemination of new items of information through exchange among members of the clusters appears to be substantiated. The system is being closely monitored for improvements of both its operations and the basic mathematical model. Work is also being

~~undertaken to expand the system to include more participants, which should lead to a larger number of more specialized clusters. Each cluster would have more members of more closely similar interests, and this should increase the probability that any member is referred to information of value to him. Hopefully, being assured to some extent that what he should know will be pointed out to him, he need not feel the necessity to read as much of the available literature as keenly as he does now.~~

ACKNOWLEDGEMENTS:

The experimental system described here is the result of the joint efforts of all members of the Information Retrieval project. The authors are particularly indebted to Mr. C. Abraham for his collaboration on analysis and design, to Mrs. H. Hoffman for collecting and collating the data and to Miss P. Reisner for many valuable contributions. Credit is also due to Mr. G. Randall and Mrs. B. Lambert of the library for serving as a message exchange, to C. Hensley, T. Tanimoto, and R. Walker for their help and last, but not least, to the participants whose initiative the effectiveness of this system is due.

NOTES:

(1) "An Operations Research Study of the Dissemination and Use of Recorded Scientific Information," Case Institute of Technology, December, 1959.

(2) G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley (1949).

(3) G. W. King and T. T. Tanimoto, unpublished.

(4) W. Brandenberg, H. C. Fallon, C. B. Hensley, T. R. Savage, and A. J. Sowarby, "Selective Dissemination of Information, SDI-2 System," IBM ASDD Technical Report 17-031.

(5) H. E. Stiles, Journal ACM, 8 , 271-279 (1961).

(6) F. Yates, Journal Royal Statistical Society, Suppl. 1 , 217-235 (1934).

(7) P. B. Baxendale, IBM Journal, 2 , 354-361 (1958).

(8) H. Cramer, *Mathematical Methods of Statistics*, Princeton University Press, 441-445 (1946).

(9) S. Chandrasekhar, Rev. Modern Physics, 15 , 1 - 89 (1943).

(10) E. W. Montroll, Nuovo Cimento, 6 Suppl. (1949).

(11) J. Neyman and E. L. Scott, Jour. Royal Stat. Soc., Series B, 20 (1958).

(12) R. D. Luce, Psychometrika, 15 , 169-191 (1950).

(13) R. D. Luce, J. Macy, Jr. and R. Taguiri, Psychometrika, 20 319-327 (1955).

(14) M. Kochen, "Organized Systems with Discrete Information Transfer," Columbia University, Appl. Math. Ph.D. Thesis, December, 1955.

(15) A. F. Parker-Rhodes and R. M. Needham, "The Theory of Clumps," Report issued by the Cambridge Language Research Unit, Cambridge, England, February, 1960.

(16) The cluster-finding problem has a number of points of contacts with other matrix permutation problems. See for example; P. C. Gilmore, "A Solution to the Module Placement Problem," IBM Research Report RC-430, (1961).

hh